NUANS minihomework 2

Fabrizio Casadei
ID: 1952529

1 Introduction

Summarization of text can be decomposed into two different problems with highly different results: **extractive summarization** and *abstractive summarization*.

In the second case, we generate a summary while rephrasing/paraphrasing the original text. This is related to *Natural language generation*. Instead in an extractive summarization, the objective is to extract relevant pieces of text from the document that have enough information to make understandable the story and contain knowledge of the crucial parts, in order to represent at best the summary of the document.

In this mini-project, it's requested to implement an *NLP* system for the extractive summarization using the provided dataset composed from a set of fairy tales and short stories, in the format of pre-numbered phrases (simple and complex) and the associated summaries from Wikipedia.

These summaries are needed for the second part of this work. indeed, a crowdsourcing process has been organized thanks to the course class, to generate the gold standard of the dataset. This process is about manually performing an extractive summarization on a small subset of the dataset's texts.

These texts have been overlapped among us students, so it's possible to first compute an evaluation from the system point of view, between the output of the model and the manually summarized text, and also between what we have personally estimated during the extractive summarization and what other students have done.

This allows a Human-Human comparison and also a Human-system comparison between the different outcomes.

2 Extractive summarization

In this section, we will briefly describe the difficulties of this task. Then it's introduced to the reader the solution proposed among different kinds of implementations conducted, using state of the art works of the *NLP* literature.

2.1 The difficulties

Extractive summarization, or in general automatic summarization tasks, are very complex activities that hide their difficulties behind the subjective nature of the operation. This complexity even increases when we pass from an extractive to an abstractive context, where an internal semantic representation is built by performing an extraction and paraphrasing/rephrasing of the original content.

In general, topic identification, interpretation, summary generation, and evaluation of the generated summary are the key challenges in text summarization.

2.2 The chosen approach

In the implementation, three different extractive summarization methods have been defined. The first uses a summarizer based on *BERT*, while the second and the third both use a variant, *S-BERT*(Reimers and Gurevych, 2019). For what concerns the difference between the second and third approaches, the second exploits the summarizer output, doing some extra operations which will be discussed later, while the third takes the embeddings from the last 3 layers, averages them, and uses these representations to compute the *cosine similarity* between the text and the summary from the dataset. This distance is then used to choose which sentence should be part of the result.

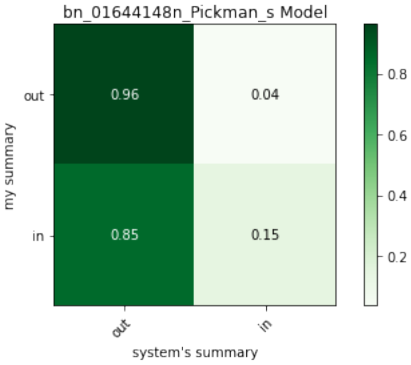
We will focus our attention on the second method mainly, which is the one used for the result.tsv file. The extractive summarizer used is based on *S-BERT* or *Sentence-BERT* (Reimers and Gurevych, 2019), which is a modification of the pre-trained *BERT* network (Devlin et al., 2018). It uses a siamese and triplet network (similarity + clustering) to derive semantically meaningful embeddings, also thanks to a pre-trained version of *RoBERTa* .(Liu et al., 2019) which has been fine-tuned properly. The resulting network presents an evident benefit for the execution time while maintaining the accuracy of *BERT*. The summarizer (Miller, 2019) internally performs its own *Sentence Boundary Detection* giving in the result, not the original sentences assigned in the dataset, but sentences by the dataset that are been cut, split and merged. The extra operations previously cited, are about transforming this outcome in order to have only the numbered phrases from the dataset. This is solved using a precise pattern that first tries to reconstruct the sentence in one to one fashion and then looks for grouping or division. This is fundamental for the integrity of the extractive task's nature.
Moreover, the model can internally decide how extended a summary can be. But has been figured out that this number is seemed insufficient to have an effective, brief and complete description of a document. Thanks to the possibility provided by the summarizer to have a degree of freedom on this number, a varying limit of sentences for each summary has been chosen. This limit has been shaped as a logarithmic function which restricts the outcome to have at max half of the original document number of sentences.

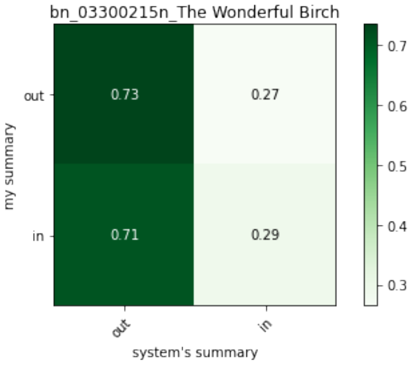
3 Human-System comparison

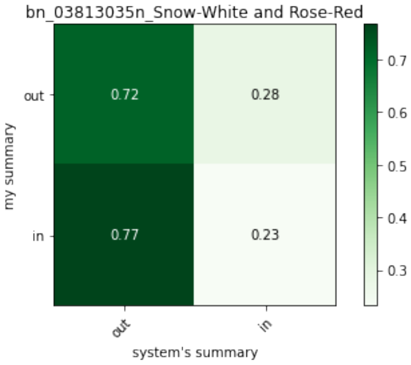
In this section, we continue the discussion on the system by giving an evaluation. The evaluation is in both quantitative and qualitative terms. Here the results are discussed in terms of the three texts assigned: *Pickman's model*, *The wonderful Birch*, *Snow-White and Rose Red*.

3.1 Quantitative analysis

How to evaluate an extractive summary as a narrative understanding system?" Here we propose an intrinsic evaluation, using as a gold standard the work manually done annotating the sentences to compose the summary.
In the following tables, that group all the metrics used for the evaluation, **S** indicates the summaries from the system, while **H** represents the ones that I have manually extracted. To notice that among the metrics chosen, mainly are present classification metrics but also *ROUNGE-L*, a standard metric for the evaluation of summarization.







	Pickman's model	The wonderful Birch	Snow-white and Rose-red
S n. sent.	16	51	27
H n. sent.	61	86	43
Accuracy	0.76	0.52	0.51
Precision	0.56	0.5	0.37
Recall	0.15	0,3	0.23
F1-score	0.23	0.36	0.28
Rouge-L P	0,76	0.76	0.57
Rouge-L R	0.3	0.47	0.44
Rouge-L F1	0.43	0.59	0.5

Instead using the *s-bert* distance method, the metrics are the ones below.

	Pickman’s model	**The wonderful Birch**	**Snow-white and Rose-red**
S n- sent.	8	33	22
H n. sent.	61	86	43
Accuracy	0.77	0.61	0.76
Precision	0.75	0.73	0.9
Recall	0.1	0.27	0.47
F1-score	0.17	0.4	0.6
Rouge-N P	0.92	0.912	0.94
Rouge-N R	0.18	0.46	0.58
Rouge-N F1	0.3	0.6	0.71

3.2 Qualitative analysis

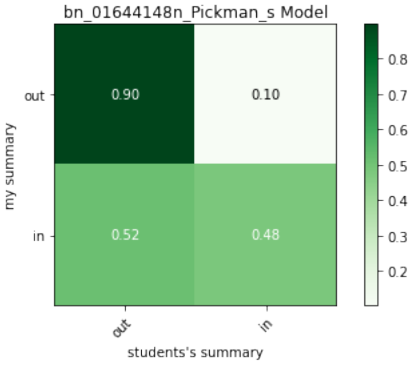
Let’s start the discussion on the qualitative difference between these two types of work. The first difference that comes up, looking at the tables above and looking at the summaries, is the length. In fact, if the automatic system tends to over-summarize the document, my manual summaries here tend to under-summarize the texts. Even trying to adjust the number of sentences to return from the system, also using a technique like *ELBOW* for clustering, the number of extracted sentences is always very reduced, therefore some key events are removed. Since the automatic system doesn’t use the *wikipidea* summary as a base for the extraction, it’s more difficult for the system to automatically detect all the key events and important concepts to insert. Moreover, now I discuss some kind of reasoning done during the manual summarization that led to this high number of extracted sentences, which aren’t taken into consideration by the automatic system.
First of all, is important to remember what are the directives to select a sentence as a part of the manual summary. If a sentence contains part of the information contained in the *wikipidea* summary, should be included. Also, we have to insert sentences to avoid inconsistency. Moreover, the final extractive summary has to be coherent and readable. The first reason why the manual summary contains so many sentences is since are needed more than one extractive sentence to describe all the information in a single reference sentence (summary from Wikipedia in the dataset), that contains multiple events and important concepts. To achieve consistency sometimes we have to include sentences to resolve pronouns and subject ambiguity as well. These problems are very present when we are facing dialogues, like for example in *Pickman’s model*, but in general is harder to summarize with a small ratio of sentences when the text has a large part composed of dialogues, even because we want a result that is readable and understandable. Another cause of this under-summarization is to include knowledge that describes an event repeated over time or a routine that is not explicitly written in the text. This, in the extractive summarization, corresponds to take different sentences, at least two but sometimes even more is necessary. Here actually is not of interest what happens in the details but this information is needed to describe a main event. This drastically reduces the summarizing effect of the work. This problem is in general extended to all key parts of the text that are not explicitly described like in the Wikipedia summary sentence. Let’s see an example, the sentence: ”As the two delved deeper into Pickman’s mind and art, the rooms seemed to grow ever more evil and the paintings ever more horrific”. Here the author refers to different scenes where there is an increase of horrific representations from Pickman’s art during the visit to his gallery. This kind of increasing gradient of obscure representations should be modeled by selecting at least two different situations in which the protagonist discusses them. These are some reasons why the manual and system summaries differ so much.

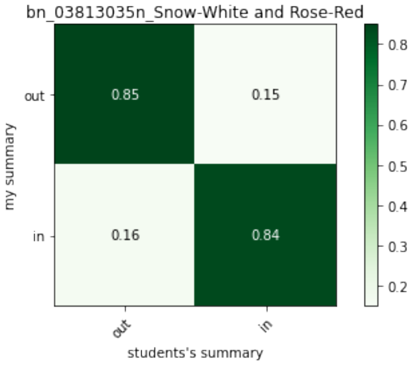
4 Human-Human comparison

After the Human-System comparison, it’s also important to understand how much the same task behaves similarly in a Human-Human context. This can be seen also as a way to establish how gold standards vary between different sources. Even there we divide the discussion into both quantitative and qualitative terms, comparing my summaries with the work done from the class. From the manual summaries performed by other students, the discussion will be in terms of the two documents: *Pickman’s model*, *Snow-White and Rose Red*. This is because for *The Wonderful Birch* no summary has been returned in the ”annotations_index.tsv” file.

4.1 Quantitative analysis

As before are presented the confusion matrices for the two documents and a table that serves as a wrap-up of the metrics utilized for the evaluation. In the table that follows, **H1** represents my manual summaries while **H2** and **H3** are the ones from other students. The values of the confusion matrix are comprehensive of both the work from H2 and H3, averaged.





	Pickman's model	**Snow-white and Rose-red**
H1 n. sentences	61	43
H2 n. sentences	52	52
H3 n. sentences	47	38
Accuracy	0.8	0.84
Precision	0.6	0.8
Recall	0.48	0.83
F1-score	0.53	0.81
Rouge-N precision	0.74	0,89
Rouge-N recall	0.58	0.86
Rouge-N f1	0.66	0,88

4.2 Qualitative analysis

Analyzing my manual annotation and the ones carried out by other students, it's possible to point out different insights.
First of all, the differences regardings the length of the summaries, and how expected are reduced with respect to the Human-system case. Although still, this difference is not neglectable. We can also suppose that, a document like *Snow-White and Rose Red* is easier to summarize respect *Pickman's model*, not only because is shorter but also because the text results simpler, in the sense that key events and concepts are found more easily and explicitly in the text. For this, the results shown in the confusion matrix and table of *Snow-White and Rose Red* show a correspondence between the three different annotations. In *Pickman's model* my annotations and the ones from other students differ a lot. As explained in the Human-System section, I had the necessity to introduce different sentences to make the summary coherent, readable, and not ambiguous. This is also part of the subjective interpretation of the work. In Fact, all the key parts of the story are present also in the shorter summaries, but they result harder to read and to fully understand. Moreover, it's important to notice the possibility that an important aspect of the story can be also described by choosing different sentences than the ones chosen in another summary.

5 Conclusion

Thanks to this short work, has been shown how is difficult to achieve good outcomes in the automatic extractive summarization task. Not only the results look different from the Human-System perspective using state of the art models, but also differ in a Human-Human comparison. This underlines the subjective nature of the task. Probably constraining the problem could lead to more convergent results for the two types of comparison performed, i.e. using a fixed number of sentences to return.
In the end, should be not neglected the typology of texts analyzed that is not the majority on which NLP tools are specialized, tendentially is more complex to interpret these kind of works.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Derek Miller. 2019. Leveraging bert for extractive text summarization on lectures. *arXiv preprint arXiv:1906.04165*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.